The Multivac: Blind Peer Matrix Evaluation of Frontier Language Models

Yash Darji
Independent Researcher
ORCID: 0009-0009-6895-842X
yashdarji2378@gmail.com

Abstract

Single-judge evaluation of large language models introduces systematic bias: GPT-4's win rate for its own outputs is approximately 10 percentage points higher than under human judging, while Claude-v1's is 25 points higher [21]. We propose the **Blind Peer Matrix**, a multi-judge evaluation methodology in which N frontier models both generate responses and evaluate each other's outputs in a fully blinded $N \times N$ matrix, with self-judgments excluded. We apply this method across 286 evaluations spanning 198 unique questions in 9 category pools (code generation, reasoning, analysis, communication, meta-alignment, edge cases, and three focused model-family batches), producing 22,254 valid judgments from 55 models. Key findings: (1) no single model dominates across all categories—six different models hold the top position across the nine pools, contradicting aggregate leaderboard rankings; (2) same-family rating bias is statistically significant in all 8 families tested ($p < 0.05$), ranging from $+0.91$ (Qwen) to -1.02 (Mistral), revealing both in-group preference and a previously unreported negative bias pattern; (3) judge disagreement is category-dependent, with code evaluation producing nearly double the disagreement of meta-alignment tasks ($\sigma = 1.27$ vs. 0.71). We release the complete evaluation dataset (27,540 judgments including self-exclusions), an open-source evaluation framework, and all question prompts under MIT license to enable reproducible, category-specific model comparison.

Keywords: LLM evaluation, peer evaluation, multi-judge, model comparison, benchmark contamination, judge disagreement

1 Introduction

How good is a language model? The question is deceptively simple. In practice, the answer depends on what you ask, how you score the response, and—critically—who does the scoring. The dominant paradigm for LLM evaluation relies on either static benchmarks or single-model judges, both of which introduce systematic distortions that can mislead practitioners, researchers, and the public. This paper proposes and validates an alternative: a blind peer evaluation methodology in which frontier models simultaneously generate responses and judge each other, producing a multi-perspective assessment that no single evaluator can provide.

1.1 The Benchmark Crisis

Static benchmarks have been the primary instrument for comparing language models since the introduction of GLUE [18] and its successors. MMLU [6], HumanEval [2], and GSM8K [3] remain widely cited as evidence of model capability. Yet these benchmarks suffer from a fundamental vulnerability: as models train on increasingly large and poorly documented corpora, the probability of test set contamination rises monotonically.

The contamination problem has moved from theoretical concern to documented practice. The retirement of the HuggingFace Open LLM Leaderboard in March 2025 was driven in part by widespread suspicion that leading models had been optimized against its test sets. More directly, Singh et al. [14] documented that Meta tested 27 private Llama-4 variants on Chatbot Arena before public release, publishing only the highest-performing result. Separately, they estimate that privileged access to Arena data can yield relative performance gains of up to 112% on the Arena distribution, demonstrating how data access asymmetries compound selective publication to distort public rankings.

Even absent deliberate gaming, static benchmarks face a ceiling problem. As frontier models approach perfect scores on MMLU and HumanEval, these benchmarks lose discriminative power. LiveBench [19] addresses this by regularly refreshing its question set, but the core limitation remains: any fixed evaluation protocol becomes a target for optimization.

1.2 The Single-Judge Problem

The LLM-as-a-Judge paradigm [21] emerged as a scalable alternative to human evaluation. MT-Bench and Chatbot Arena demonstrated that GPT-4 judgments correlate reasonably with human preferences, establishing LLM-as-a-Judge as a practical evaluation tool.

However, the same study that validated the approach also documented its failure modes. Three biases are well-established:

Self-enhancement bias. Zheng et al. [21] observe that GPT-4's win rate for its own outputs is approximately 10 percentage points higher than under human judging, and Claude-v1's approximately 25 points higher, though they note that the available data are insufficient to definitively establish self-enhancement bias. Subsequent work has confirmed the pattern at larger scale: models systematically rate outputs from their own family more favorably, suggesting they internalize quality criteria correlated with their own generation patterns.

Position bias. Responses presented first in a pairwise comparison are rated higher regardless of quality. Zheng et al. [21] found this effect to be statistically significant across evaluation categories.

Verbosity bias. Longer responses receive higher scores independent of content quality. This creates a perverse incentive: models optimized for evaluation performance learn to produce verbose outputs that score well on surface metrics while adding little substantive value. The clarity–depth gap we document in §5.2—where models score 0.8–1.4 points higher on clarity than on depth—may partially reflect this dynamic.

The Panel of LLM Judges (PoLL) approach [17] demonstrated that using multiple judge models reduces these biases and correlates better with human judgments than any single judge. Our work extends this insight into a fully symmetric framework where every model serves as both contestant and judge.

1.3 Our Contribution

We make three contributions:

1. The Blind Peer Matrix methodology. We formalize an $N \times N$ multi-judge evaluation protocol in which N frontier models generate responses to a shared prompt and then evaluate each other's outputs under blinded conditions. Self-judgments are excluded. For a pool of $N = 10$ models, each evaluation produces 90 independent judgments scored across five weighted dimensions. The methodology is fully specified in §3 and implemented as open-source software.

2. Category-specific evaluation at scale. We apply the Blind Peer Matrix across 286 evaluations spanning 198 unique questions in 9 category pools, producing 22,254 valid judgments from 55 models. We demonstrate that aggregate rankings mask substantial per-category variation: six different models hold the top position across the nine pools. We report the first large-scale analysis of same-family rating bias in multi-judge evaluation, finding statistically significant bias in all 8 families tested, ranging from $+0.91$ (Qwen) to -1.02 (Mistral).

3. An open disagreement dataset. We release the complete evaluation dataset—27,540 judgments (22,254 valid, 5,286 self-excluded), all question prompts, model responses, per-dimension scores, and judge identities—under MIT license. To our knowledge, this is the largest publicly available multi-judge LLM evaluation dataset with full provenance.

2 Related Work

We situate our work at the intersection of three lines of research: LLM evaluation frameworks, multi-judge approaches, and the emerging study of systematic disagreement between models.

2.1 LLM Evaluation Frameworks

Static benchmarks. The dominant paradigm evaluates models against fixed test sets. MMLU [6] tests factual knowledge across 57 domains; HumanEval [2] and MBPP [1] evaluate code generation; GSM8K [3] targets mathematical reasoning. HELM [11] aggregated multiple benchmarks into a unified evaluation harness. These benchmarks enabled rapid progress but share a structural vulnerability: once published, they become optimization targets.

LLM-as-a-Judge. Zheng et al. [21] introduced MT-Bench and Chatbot Arena, demonstrating over 80% agreement with human preferences. G-Eval [12] extended the approach with chain-of-thought scoring. Prometheus 2 [7] trained a dedicated evaluation model to follow rubrics.

Dynamic benchmarks. LiveBench [19] addresses contamination by regularly refreshing its question set. Arena-Hard [10] curates difficult questions from Chatbot Arena. Both retain single-judge scoring, inheriting the biases documented in §1.2.

Leaderboard gaming. Singh et al. [14] provided direct evidence that evaluation gaming extends beyond contamination, documenting selective publication from privately tested model pools.

2.2 Multi-Judge Approaches

Verga et al. [17] introduced the Panel of LLM Judges (PoLL), demonstrating that a panel of diverse models reduces intra-model bias and correlates better with human judgments than any single judge. Our Blind Peer Matrix extends PoLL in two directions: full symmetry (every model is both contestant and judge) and scale (90 judgments per evaluation from 10 simultaneous judges).

JudgeBench [15] creates a benchmark for evaluating evaluator models. Two comprehensive surveys [9, 5] catalogue the growing LLM-as-a-Judge literature, identifying single-judge bias as the central unresolved problem.

2.3 Disagreement and Uncertainty

A critical question for any multi-judge system is whether inter-judge disagreement is noise or signal. Shi et al. [13] provided evidence for the latter in a large-scale study of 15 LLM judges across over 150,000 evaluation instances. They found full unanimity in only approximately 23% of cases, with models sharing architecture and training lineage exhibiting systematically higher internal agreement—a family clustering effect that suggests disagreement reflects genuine differences in evaluative criteria rather than random noise.

Our disagreement analysis (§5.4) extends this by documenting that disagreement varies systematically by task category. The broader uncertainty quantification literature [4, 20] has explored single-model confidence, but our peer matrix provides an external, multi-perspective measure that does not depend on any individual model's calibration.

2.4 Positioning

Three properties distinguish our approach: (1) full symmetry—every model is both judge and contestant; (2) scale—90 judgments per evaluation from 10 simultaneous judges; and (3) transparency—we release all judgments, scores, and prompts.

3 Methodology

3.1 Formal Definition

Let $M = \{m_1, m_2, \ldots, m_N\}$ be a set of N language models and let q be a question prompt drawn from a category-specific question bank Q.

Response Generation. Each model $m_i \in M$ generates a response $r_i = m_i(q)$. All N models receive the identical prompt q with no system-prompt variation.

Judgment. Each model $m_j \in M$ evaluates each response r_i where $i \neq j$, producing a score vector:

$$\mathbf{s}_{ij} = m_j(r_i, C) = (\text{correctness}_{ij},\ \text{completeness}_{ij},\ \text{clarity}_{ij},\ \text{depth}_{ij},\ \text{usefulness}_{ij})$$

The constraint $i \neq j$ enforces **self-exclusion**. This produces an $N \times N$ score matrix S where entry $S_{ij} = \mathbf{w} \cdot \mathbf{s}_{ij}$ is the weighted composite score, and the diagonal is empty. For $N = 10$, each evaluation produces 90 valid judgments.

Aggregation. The final score for respondent m_i on question q is:

$$\text{score}(m_i, q) = \frac{1}{N-1} \sum_{j \neq i} S_{ij}$$

3.2 Scoring Rubric

Judges evaluate each response along five dimensions using a 0–10 scale. The composite weighted score is:

$$S_{ij} = 0.25 \cdot \text{correctness}_{ij} + 0.20 \cdot \text{completeness}_{ij} + 0.20 \cdot \text{clarity}_{ij} + 0.20 \cdot \text{depth}_{ij} + 0.15 \cdot \text{usefulness}_{ij}$$

Correctness receives the highest weight; usefulness the lowest due to susceptibility to verbosity bias. The rubric is provided to judges as a structured system prompt specifying each dimension's definition and weight. The exact judge prompt is reproduced in Appendix D.

3.3 Blinding and Anti-Contamination

Blinding protocol. Three measures ensure judges cannot identify which model produced a response: (1) *Identity masking*—judges receive only raw response text; (2) *Order randomization*—response order is randomized per judge per evaluation; (3) *Low-temperature judging*—response generation uses temperature 0.7 (default) to preserve natural diversity, while judge API calls use temperature 0.3 to reduce stochastic variation in scoring while avoiding edge-case behavior some providers exhibit at temperature 0.

Anti-contamination. All 198 questions are written de novo (no existing benchmarks), designed to require cross-domain synthesis, and used at most once. The question bank grew from 60 to 198 questions over the evaluation period.

3.4 Category-Specific Model Pools

We compose category-specific pools of $N = 10$ models: six primary categories (code, reasoning, analysis, communication, meta-alignment, edge cases) and three focused batches (SLM, Qwen, MiniMax). Pool composition is detailed in §4.1.

3.5 Head-to-Head Protocol

As a complement, we implement head-to-head evaluation with a single external judge (Claude Opus 4.6) for rapid pairwise comparison. H2H results are reported separately (§5.6) as validation.

3.6 Implementation

The evaluation framework is implemented in Python, orchestrating API calls via OpenRouter and direct provider APIs. Results are stored as structured JSON with full provenance. The complete pipeline, dataset (27,540 judgments), all 198 question prompts, and scoring rubric are released under MIT license.[1]

[1] Code and dataset: https://github.com/themultivac/multivac-evaluation. Platform: https://app.themultivac.com.

4 Experimental Setup

4.1 Models

A total of 55 distinct models participated. The core pool of 10 frontier models (GPT-5.4, Claude Opus 4.6, Claude Sonnet 4.6, GPT-OSS-120B, MiMo-V2-Flash, Grok 4.20, DeepSeek V3, Gemini 3.1 Pro, Gemini 3 Flash Preview, MiniMax M2.5) appears in 116–230 evaluations each. An extended pool of 11 models participates in 30–99 evaluations, and 34 focused batch models appear in 1–29 evaluations. The full model list is in Appendix A.

4.2 Question Design

All 198 questions were authored de novo following four principles: discriminative power (targeting the capability frontier), cross-domain synthesis (reducing memorization advantage), category alignment (one category per question), and non-reuse (each question appears once). Distribution: code (50), analysis (49), communication (48), reasoning (47), meta-alignment (44), SLM (14), MiniMax (13), Qwen (11), edge cases (10). Sample questions are in Appendix B.

4.3 Evaluation Protocol

Evaluations were conducted February–April 2026 across two phases. Phase 1 (February–March) iteratively refined the scoring rubric, judgment prompt, and parsing logic, reducing parse failure from 41.5% to approximately 10%. Two Phase 1 judgments containing scores of 100 (on a 0–10 scale) due to absent range clamping were identified and excluded. Phase 2 (March–April) applied the mature protocol at scale. See Appendix D for the full Phase 1/Phase 2 protocol comparison.

Cost. Each peer matrix evaluation costs $2–3 in API fees. The complete dataset cost approximately $700–850.

4.4 Judgment Parsing

The parsing pipeline extracts per-dimension scores using a multi-strategy approach: thinking-block removal, JSON brace matching, regex fallback, and 0–10 range clamping. Of 27,540 total judgments, 22,254 (80.8%) produced valid scores after self-exclusion.

5 Results

We report results across 286 peer matrix evaluations spanning 198 unique questions and 9 category pools, yielding 22,254 valid judgments from 55 models. Self-judgments ($N = 5{,}286$) were excluded per protocol.

5.1 Overall Cross-Category Rankings

Several patterns are notable. First, the model with the most first-place finishes is not the highest-scoring model by mean. GPT-5.4 accumulates 53 wins across 186 evaluations yet its mean score (8.946) places it 16th overall. By contrast, Seed 1.6 Flash achieves the highest mean (9.425) on only 10 evaluations.

Second, models with high evaluation counts cluster in a narrow band. The six models appearing in 180+ evaluations span a range of only 0.477 points, despite representing five different model families. This compression at scale suggests frontier models converge in aggregate quality while diverging on specific task types.

Statistical significance of rankings. We compute bootstrap confidence intervals (5,000 iterations) for each model with ≥ 30 evaluations ($N = 21$ models). The top 9 models—from Grok 4.1 Fast (9.394, 95% CI [9.052, 9.863]) to Gemini 2.5 Flash (8.873, 95% CI [8.295, 9.311])—have substantially overlapping confidence intervals. Of 45 pairwise comparisons among the top 10, only 7

reach significance at $p < 0.05$. The top 4 models are pairwise indistinguishable (all $p > 0.07$). For the top tier of frontier models, aggregate ranking differences are not statistically meaningful.

Below the top tier, separation is clearer. GPT-5.4 (ranked 7th) is significantly better than Gemini 3 Flash Preview (ranked 10th, $p = 0.002$). Gemini 3.1 Pro (ranked 21st, 6.449, 95% CI [6.121, 6.780]) is significantly below every model ranked 18th or higher.

5.2 Category-Specific Analysis

Six different models hold the top position across the nine pools. Claude Sonnet 4.5 leads analysis (9.615) and communication (9.568) but drops to 7th in code. GPT-5.4 ranks 1st in none of the six primary categories despite leading in overall wins. Grok-family models dominate code and edge cases but are absent from the top-3 in analysis.

Small Language Models. Within the SLM pool ($\leq$32B parameters), Qwen 3 8B (9.328) outperforms models with 2–4$\times$ its parameter count, including Phi-4 14B (8.924) and Gemma 3 27B (8.853).

Dimension-level analysis. Across all models, depth is consistently the lowest-scoring dimension (mean: 8.45) while clarity is the highest (mean: 9.40). The clarity–depth gap exceeds 0.8 points for 17 of 55 models, with GPT-5.2-Codex showing the largest gap ($\Delta = 1.438$). This suggests current frontier models are better calibrated for well-structured output than for deep analytical engagement.

5.3 Judge Behavior Analysis

Judge behavior varies dramatically. The strictest high-volume judge is GPT-5.4 (mean 7.187 across 1,565 judgments); the most lenient is Mistral Small Creative (9.695 across 422 judgments)—a spread of 2.51 points. OpenAI models exhibit high variance as judges (GPT-5.4 std = 2.215), while Seed 1.6 Flash (std = 0.877) and Granite 4.0 Micro (std = 0.488) compress scores into a narrow band.

5.4 Disagreement Patterns

Code evaluation produces the highest average disagreement ($\sigma = 1.269$), nearly double that of meta-alignment ($\sigma = 0.705$). Edge cases show high mean disagreement (1.248) but notably lower median (0.670), indicating a skewed distribution with occasional extreme disagreements. The gap between mean and median across all categories indicates right-skewed distributions: most evaluations produce moderate agreement, but a tail of high-disagreement cases exists in every category.

5.5 Same-Family Rating Bias

We map each of the 55 models to one of 11 developer families and compute bootstrap confidence intervals (10,000 iterations) for the bias estimate. All eight family-level bias estimates with sufficient data are statistically significant ($p < 0.05$), and seven survive Bonferroni correction ($\alpha = 0.00625$).

Table 1: Same-family rating bias with bootstrap 95% confidence intervals.

Family	Bias	95% CI	p-value	N_{same}	N_{other}
Qwen	$+0.913$	$[+0.603, +1.234]$	< 0.0001	434	130
xAI	$+0.745$	$[+0.405, +1.012]$	0.0001	59	2,269
Anthropic	$+0.616$	$[+0.486, +0.740]$	< 0.0001	482	3,719
MiniMax	$+0.314$	$[+0.077, +0.543]$	0.005	245	1,493
OpenAI	$+0.229$	$[+0.033, +0.420]$	0.011	385	3,315
Google	-0.593	$[-0.817, -0.379]$	< 0.0001	426	3,346
Meta	-0.681	$[-1.361, -0.200]$	0.0008	26	121
Mistral	-1.017	$[-1.359, -0.693]$	< 0.0001	25	532

Five families exhibit positive bias (rating siblings higher); three exhibit negative bias. The strongest positive bias belongs to Qwen ($+0.913$), meaning Qwen-family judges rate other Qwen

models nearly a full point higher on a 10-point scale. The negative biases are equally notable: Mistral models rate siblings 1.017 points *lower* than non-Mistral models—the largest absolute bias. This negative pattern is, to our knowledge, previously unreported.

5.6 Head-to-Head Validation

Five H2H batches (180 total questions) using Claude Opus 4.6 as single judge validate the peer matrix results. The largest batch (Qwen 3.6 Plus vs. DeepSeek V3.2, 150 questions) showed Qwen winning 107–11 with 25 ties. H2H batches serve as rapid signals subject to single-judge biases (§5.3).

6 Limitations

6.1 LLMs Evaluating LLMs

Multi-judge aggregation reduces individual biases but does not eliminate biases shared across all judges. The clarity–depth gap (§5.2) may reflect exactly this kind of shared bias. A human correlation study is needed.

6.2 Output Quality vs. Reasoning Process

The methodology evaluates final output quality without examining the reasoning process. A model that arrives at a correct answer through flawed reasoning receives the same score as one demonstrating genuine understanding. Future work should incorporate reasoning trace evaluation [16].

6.3 Sample Size and Statistical Power

Per-category and per-model sample sizes vary considerably. Core pool models participate in 116–230 evaluations; focused batch models may appear in as few as 7–10. Bootstrap confidence intervals (§5.1) quantify ranking uncertainty for the 21 models with ≥ 30 evaluations.

6.4 Question Design Bias

The first author wrote all 198 questions, introducing unavoidable perspective bias. We mitigate through category diversity, volume, and fresh daily authorship.

6.5 Cost and Throughput

At $2–3 per evaluation, the methodology is expensive relative to static benchmarks but inexpensive relative to human evaluation.

6.6 Inter-Annotator Agreement

We compute Krippendorff's α (interval level) across all judge–respondent pairs. The overall $\alpha = 0.618$, approaching the conventional threshold of 0.667 for tentative conclusions [8]. Two categories—reasoning ($\alpha = 0.687$) and MiniMax ($\alpha = 0.673$)—exceed this threshold, while communication ($\alpha = 0.354$) and edge cases ($\alpha = 0.436$) show lower agreement. The reasoning result is notable: judges converge more readily on logical analysis quality than on communication quality, consistent with the intuition that reasoning has more objective markers.

The moderate overall agreement supports our core methodological argument: single-judge evaluation is unreliable precisely because judges disagree systematically. An α of 0.618 indicates judges share enough common ground to produce meaningful aggregate rankings while disagreeing enough that no single judge's perspective dominates.

7 Discussion

7.1 Implications for Model Selection

Model selection should be task-specific. The finding that GPT-5.4 leads in wins (53) but ranks 16th by mean score illustrates that win-count and mean-score rankings answer different questions. Category-specific rankings (§5.2) provide more actionable signal than any aggregate.

7.2 Implications for AI Safety

The same-family bias results (§5.5) imply that single-family evaluation of alignment properties may produce misleadingly optimistic results. The low disagreement on meta-alignment ($\sigma = 0.705$) may reflect shared RLHF biases rather than genuine consensus.

7.3 The Data Engine

Each evaluation generates 90 preference pairs usable for DPO/RLHF training at <\$0.01/sample. The five-dimensional scoring enables optimization for specific dimensions rather than monolithic preference.

7.4 Toward Open Evaluation Infrastructure

The EU AI Act creates demand for independent evaluation. We release the evaluation framework, question bank, scoring rubric, and complete dataset under MIT license. At current API prices the full matrix costs \$2–3 per evaluation; we are exploring open-source judge distillation to reduce inference costs by an estimated 70%, which would make continuous evaluation economically viable at scale.

8 Conclusion

We introduced the Blind Peer Matrix, a multi-judge evaluation methodology in which N frontier language models simultaneously generate responses and evaluate each other's outputs under blinded conditions. Applied across 286 evaluations spanning 198 unique questions and 9 category pools, the method produced 22,254 valid judgments from 55 models—to our knowledge, the largest publicly available multi-judge LLM evaluation dataset with full provenance.

Three findings stand out. First, no single model dominates across all categories: six different models hold the top position across the nine pools, and the top 4 models are statistically indistinguishable in aggregate ($p > 0.07$ pairwise). Second, same-family rating bias is statistically significant in all 8 families tested, ranging from $+0.91$ (Qwen, $p < 0.0001$) to -1.02 (Mistral, $p < 0.0001$), revealing both in-group preference and a previously unreported negative bias pattern. Third, judge disagreement is category-dependent ($\sigma = 1.27$ for code vs. 0.71 for meta-alignment), with overall inter-annotator agreement of $\alpha = 0.618$ (Krippendorff's alpha).

We release the complete dataset (27,540 judgments), all 198 question prompts, the evaluation framework source code, and documentation under MIT license.

References

[1] Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., and Sutskever, I. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[2] Chen, M., Tworek, J., Jun, H., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[3] Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059, 2016.

[5] Gu, Z., Zhu, X., Ye, H., et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2412.05579*, 2024.

[6] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *ICLR*, 2021.

[7] Kim, S., Shin, J., Cho, Y., et al. Prometheus 2: An open source language model specialized in evaluating other language models. In *EMNLP*, 2024.

[8] Krippendorff, K. Computing Krippendorff's alpha-reliability. Departmental Papers (ASC), University of Pennsylvania, 2011.

[9] Li, J., Sun, A., Han, J., and Li, C. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[10] Li, T., Chiang, W.-L., Frick, E., et al. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

[11] Liang, P., Bommasani, R., Lee, T., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[12] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *EMNLP*, 2023.

[13] Shi, L., Muennighoff, N., Liu, H., et al. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024.

[14] Singh, S., Nan, Y., Wang, A., et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025. Accepted at NeurIPS 2025 Datasets and Benchmarks Track.

[15] Tan, H., Liu, S., Su, D., et al. JudgeBench: A benchmark for evaluating LLM-based judges. *arXiv preprint arXiv:2410.12784*, 2024.

[16] Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

[17] Verga, P., Hofstatter, S., Althammer, S., et al. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[18] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

[19] White, C., Dooley, S., Roberts, M., et al. LiveBench: A challenging, contamination-free LLM benchmark. In *ICLR*, 2025.

[20] Xiao, Y. and Wang, W. Y. Quantifying uncertainties in natural language processing tasks. In *AAAI*, 2019.

[21] Zheng, L., Chiang, W.-L., Sheng, Y., et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

A Complete Model List

See supplementary materials for the full table of all 55 models with family, parameter count, evaluation count, judgments given, mean score, and category participation. Models span 11 developer families: Anthropic, OpenAI, Google, xAI, DeepSeek, Xiaomi, MiniMax, Qwen/Alibaba, Mistral, ByteDance, and Meta.

B Question Bank (Sample)

Twelve representative questions from the 198-question bank are provided in supplementary materials, covering all six primary categories with a range of difficulties. The full question bank is released with the dataset.

C Score Matrix Example

A complete 10×10 score matrix from evaluation EVAL-20260403-112809 (meta-alignment) is provided in supplementary materials, demonstrating judge strictness variation, per-respondent disagreement, and self-exclusion.

D Judge System Prompt and Protocol Evolution

The exact judge system prompt and a detailed comparison of Phase 1 vs. Phase 2 protocol changes (prompt structure, parser robustness, score validation, ranking thresholds) are provided in supplementary materials.

E Per-Model Dimension Breakdown

Mean scores by dimension (correctness, completeness, clarity, depth, usefulness) for all 30 models with $N \geq 100$ judgment observations are provided in supplementary materials.